ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΕΑΡΙΝΟ ΕΞΑΜΗΝΟ 2017-18

M112: Διαχείριση Μεγάλων Δεδομένων

Analyze customer transactions: build a system with kafka and spark to analyze customer transactions.

Κουκούλη Μυρτώ M1503 - mkoukouli@di.uoa.gr
Γαλούνη Κωνσταντίνα M1524 - kgalouni@di.uoa.gr
Κανακάκης Παναγιώτης M1516 - pkanakakis@di.uoa.gr

ΑΘΗΝΑ

Απρίλιος 2018

Πίνακας Περιεχομένων

Δεδομένα

Τα δεδομένα που χρησιμοποιούνται είναι διαθέσιμα στη διαδικτυακή πλατφόρμα για την κοινότητα των τεχνικών εξόρυξης δεδομένων, Kaggle, και είναι προσβάσιμα από τον σύνδεσμο https://www.kaggle.com/retailrocket/ecommerce-dataset. Το μέγεθός τους είναι περίπου 1 GB συνολικά.
Πρόκειται για περιγραφή, των βασικών ενεργειών των πελατών και επισκεπτών ενός ηλεκτρονικού καταστήματος, των χαρακτηριστικών των προϊόντων και των σχέσεων μεταξύ των κατηγοριών στις οποίες τα προϊόντα ανήκουν. Οι ενέργειες ενός χρήστη περιορίζονται σε προβολή ενός προϊόντος, προσθήκη στο καλάθι και αγορά, ενώ τα χαρακτηριστικά των προϊόντων αφορούν δεδομένες χρονικές στιγμές. Για την περιγραφή των παραπάνω, δίνονται 3 αρχεία csv, ένα για κάθε περίπτωση.

Στόχος Εργασίας

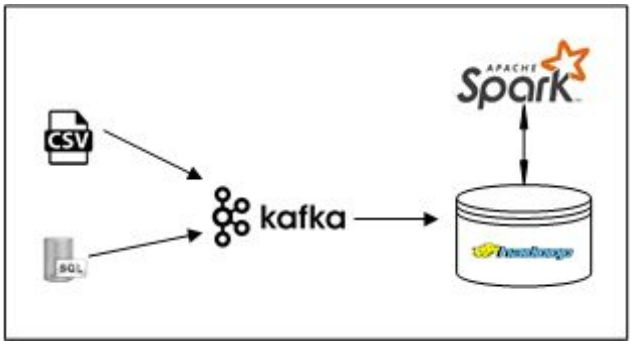
Η εργασία αυτή εκπονείται στο πλαίσιο του μαθήματος “Διαχείριση Μεγάλων Δεδομένων”. Τα δεδομένα μας οριακά (δεδομένου ότι υπάρχει περιορισμός στους διαθέσιμους πόρους) μπορούν να θεωρηθούν αρκετά μεγάλα, ώστε να εκμεταλλευτούμε τα εργαλεία και τις τεχνικές που είναι διαθέσιμες σχετικά με εκμετάλλευση πολλαπλών πυρήνων και μηχανημάτων ως clusters, για ταχύτερη διαχείρισή τους.

Για την επεξεργασία των δεδομένων θα ακολουθήσουμε την εξής σειρά:

- Ανάλυση δεδομένων: Εκτέλεση επερωτημάτων, εξαγωγή στατιστικών αποτελεσμάτων και οπτικοποίηση μέρους αυτών χρησιμοποιώντας το Spark (SparkSQL, SparkR).
- Χρήση της MLib βιβλιοθήκης του Spark για κατηγοριοποίηση των δεδομένων. Βασικός στόχος είναι η εύρεση μιας μεθόδου πρόβλεψης των μελλοντικών ενεργειών των χρήστη μέσω ανάλυσης των χαρακτηριστικών των προϊόντων για τα οποία έδειξε ενδιαφέρον.

Για την εξοικείωσή μας με περισσότερα από τα διαθέσιμα εργαλεία διαχείρισης Μεγάλων Δεδομένων, παρόλο που τα δεδομένα δεν δίνονται σε συνεχή ροή (streaming), αλλά είναι στατικά αποθηκευμένα σε αρχεία, στοχεύουμε να τα διαβάσουμε μέσω του συστήματος Kafka. Μεταξύ του Kafka και του Spark θα παρεμβάλλεται ένα HDFS.

Το πλάνο υλοποίησης απεικονίζεται στο παρακάτω σχήμα:



Εγκατάσταση εργαλείων

Για τις ανάγκες της εργασίας χρησιμοποιούμε: scala 2.11.12, Spark 2.3.0 (μαζί με Hadoop 2.7) και Kafka.

Τα εργαλεία έχουν εγκατασταθεί τοπικά.
Προς το παρόν, δεν έχουμε στήσει cluster μηχανημάτων, συνεπώς η εφαρμογή για το Spark εκτελείται τοπικά ως master. Το SparkSession είναι υπεύθυνο να εκκινήσει και να τερματίσει έναν master, εκμεταλλευόμενο όλους τους διαθέσιμους πυρήνες του μηχανήματος.

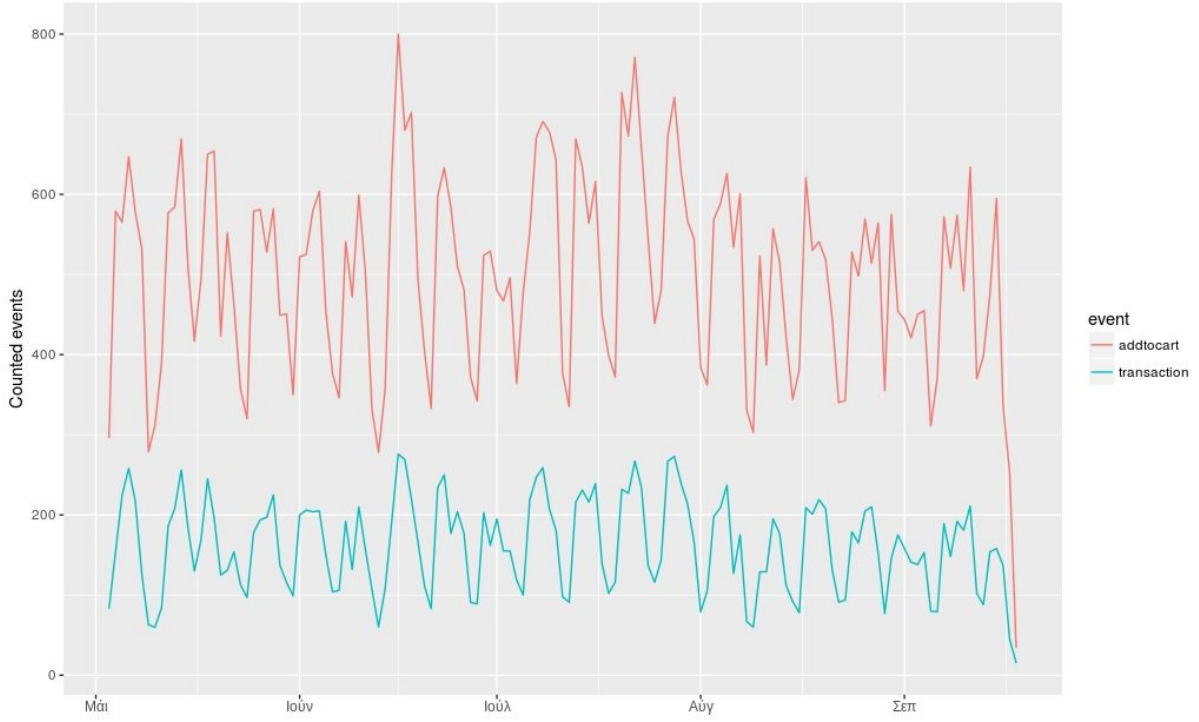
Πρόοδος εργασιών και Αποτελέσματα

Φάση 1 (Ανάλυση Δεδομένων και Εξαγωγή Στατιστικών)

Προς το παρόν έχουν εξαχθεί μερικά στατιστικά για τα δεδομένα.
Τα δεδομένα περιγράφουν τις κινήσεις χρηστών για ένα διάστημα 5 μηνών.

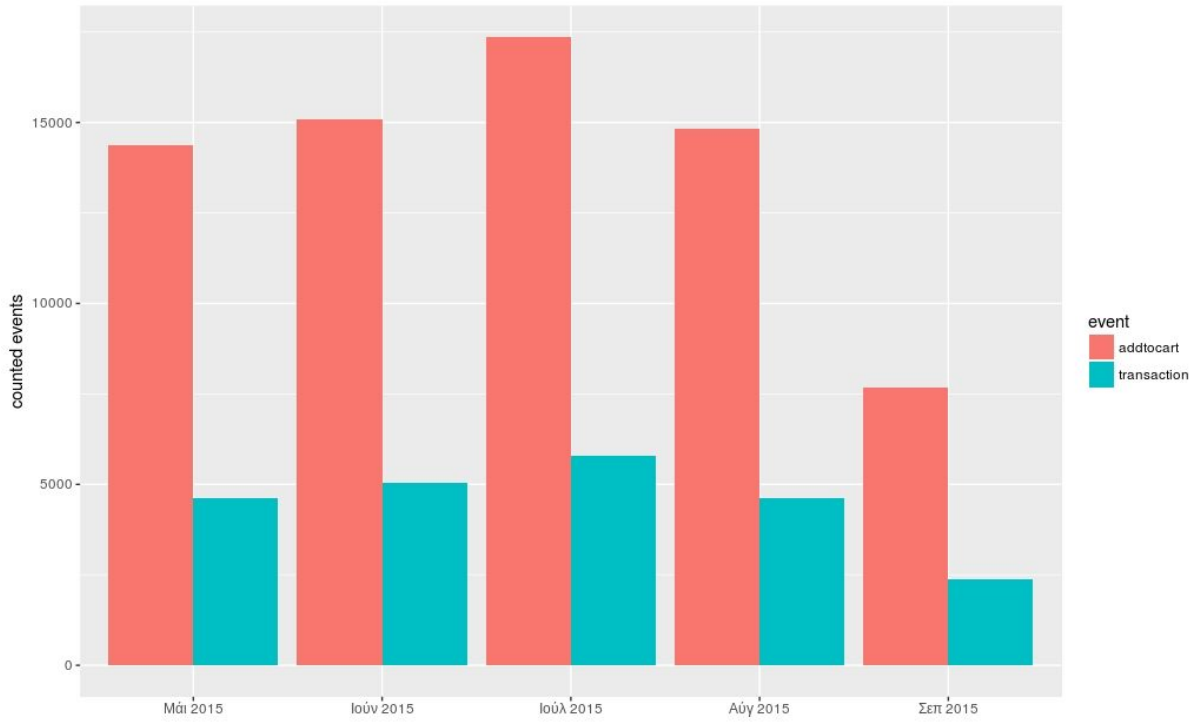
Ενδεικτικά, παρατίθενται 3 διαγράμματα σχετικά με την ανάλυση των κινήσεων των χρηστών [1].

1. Σύγκριση ημερήσιου αριθμού transactions με τα addToCart.

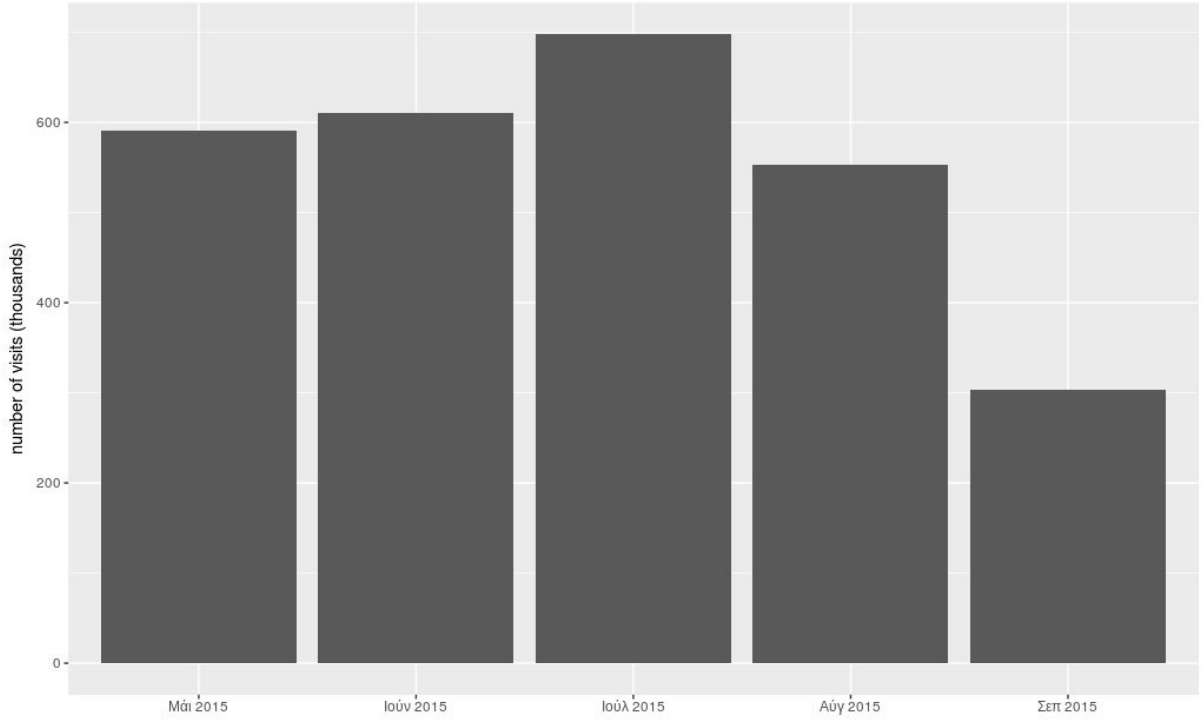


[1] Οι κινήσεις περιγράφονται μέσω καταγραφής event τύπου ‘view',’addToCart’ και ‘transaction’ και αφορούν ένα προϊόν κάθε φορά.

2. Σύγκριση μηνιαίου αριθμού transactions με τα addToCart



3. Συνολικός αριθμός καταγεγραμμένων event ανά μήνα.



Φάση 2 (Κατηγοριοποίηση - Recommendations)

Θα προχωρήσουμε στη φάση αυτή αμέσως μετά την ολοκλήρωση της φάσης 1.

Φάση 3 (Είσοδος δεδομένων από το Kafka σε Hdfs και σύνδεση με Spark)

Προς το παρόν έχει υλοποιηθεί απλή μέθοδος ανάγνωσης δεδομένων από CSV αρχεία.
Η σύνδεση με το Kafka θα υλοποιηθεί σε πιο προχωρημένο στάδιο της εργασίας.